CHAPTER 1
General introduction: evaluation of diagnostic procedures

J. André Knottnerus, Frank Buntinx, and Chris van Weel

Summary box

- Whereas the development of diagnostic technologies has greatly accelerated, the methodology of diagnostic research still lags behind that of evaluation of treatment.
- Diagnostic research appears to be more comprehensive and complex than treatment research as it evaluates the connection between diagnostic and prognostic assessment with choosing optimal interventions.
- Objectives of diagnostic testing are (1) detecting or excluding disorders, (2) contributing to further diagnostic and therapeutic management, (3) assessing prognosis, (4) monitoring clinical course, and (5) measuring general health or fitness.
- Methodological challenges include dealing with complex relations, the "gold standard" problem, spectrum and selection bias, "soft" outcome measures, observer variability and bias, optimizing clinical relevance, appropriate sample size, and rapid progress of applicable knowledge over time.
- Choosing the appropriate study design depends on the research question; the most important designs are the cross-sectional study (to determine the accuracy and added discriminatory value of diagnostic procedures) and the randomized controlled trial (to evaluate the clinical impact of [additional] testing).
- To synthesize the results of various studies on the same topic, diagnostic systematic reviews and meta-analyses are powerful tools.

(*continued*)

The Evidence Base of Clinical Diagnosis: Theory and Methods of Diagnostic Research. 2nd edition. Edited by J. André Knottnerus and Frank Buntinx. © 2009 Blackwell Publishing, ISBN: 978-1-4051-5787-2.

(*continued*)
- To make the step from research to practice, clinical decision analysis, cost-effectiveness studies, and quality-of-care research, including implementation studies, are important.

Introduction

The development and introduction of new diagnostic technologies have accelerated greatly over the past few decades. This is reflected in a substantial expansion of research on diagnostic tests and of publications on diagnostic research methodology. However, the evaluation of diagnostic techniques is far from being as established as the evaluation of therapies.

Apart from regulations for obvious safety aspects, at present—unlike the situation with regard to drugs—there are no widely accepted formal requirements for validity and effectiveness that diagnostic tests must meet to be accepted or retained as a routine part of health care. This is related to another point: in spite of important early initiatives[1,2] the methodology for evaluation of diagnostics is not yet as crystallized as the deeply rooted principles of the randomized controlled trial on therapeutic effectiveness[1,3] and of etiologic study designs.[4,5] It is not surprising, then, that serious methodological flaws are often found in published diagnostic studies.[6,7,8]

An additional challenge is the comprehensiveness of diagnostic research. Diagnostic evaluation is the first crucial medical intervention in an episode of illness, labeling symptoms and complaints as health problems, and indicating possible disease and its prognosis. Effective and efficient therapy—also including reassurance, "watchful waiting," and supporting patient self-efficacy—depends to a large extent on an accurate interpretation of (early) symptoms and the outcome of the diagnostic process. Accordingly, as diagnosis covers not only test accuracy but is also the basis for prognosis and appropriate treatment choice, diagnostic research is in fact much more complex than treatment research.

A special point of concern is that the funding of diagnostic evaluation studies is not well organized, especially if the research is not focused on particular body systems or disorders covered by strong research foundations. Rather than being limited to a particular body system, diagnostic evaluation studies frequently start from a complaint, a clinical problem, or certain tests.

Because the quality of diagnostic procedures is indicative for the quality of health care as a whole, it is vital to overcome the shortfall in standards, methodology, and funding. Accurate evaluation of diagnostic performance will contribute to the prevention of unjustified treatment, lack of treatment or mistreatment, and unnecessary costs. In this context, important steps forward toward the professionalizing of diagnostic studies have already been made with the work on the architecture of diagnostic studies[9] (see also Chapters 1, 2, and 14), the Standards for Reporting of Diagnostic Accuracy studies (STARD)[10]

(see Chapter 9), and QUADAS, a tool for the Quality Assessment of Diagnostic Accuracy Studies included in systematic reviews.[11] As a general background, this introductory chapter presents an overview of the objectives of diagnostic testing and evaluation research, important methodological challenges, and research design options.

Objectives

Diagnostic testing can be seen as the collection of additional information intended to (further) clarify the character and prognosis of the patient's condition and can include patients' characteristics, symptoms and signs, history and physical examination items, or additional tests using laboratory or other technical facilities. A "test" not only must be considered but also the specific question the test is supposed to answer. Therefore, the performance of tests must be evaluated in accordance with their intended objectives. Objectives may include:

- *Detecting or excluding disorders, by increasing diagnostic certainty as to their presence or absence.* This can only be achieved if the test has sufficient discrimination. Table 1.1 shows the most common measures of discrimination. Most of these can be simply derived from a 2×2 table comparing the test result with the diagnostic reference standard, as demonstrated by the example of ankle trauma. A more elaborate and comprehensive explanation of how to calculate these and other measures from collected data is presented in Chapter 7. Examples of tests for which such measures have been assessed are given in Table 1.2. Such a representation allows various tests for the same purpose to be compared. This can show, for example, that less invasive tests (such as ultrasonography) may be as good as or even better diagnostically than more invasive or hazardous ones (e.g., angiography). Also, it can be shown that history data (e.g., change in bowel habit) may be at least as valuable as laboratory data. What is important is not just the discrimination per se, but rather what a test may add to what cheaper and less invasive diagnostics already provided to the diagnostic process. This is relevant, for instance, in assessing the added value of liver function tests to history taking and physical examination in ill-defined, nonspecific complaints. Finally, using the measures defined in Table 1.1 allows the comparison of the value of a test in different settings, for example, general practice versus the hospital emergency department.[12]
- *Contributing to the decision-making process with regard to further diagnostic and therapeutic management*, including the indications for therapy (e.g., by determining the localization and shape of a lesion) and choosing the preferred therapeutic approach.
- *Assessing prognosis* on the basis of the nature and severity of diagnostic findings. This is a starting point for planning the clinical follow-up and for informing and—if justified—reassuring the patient

Table 1.1 Commonly used measures of the discrimination of a diagnostic test T for disease D, illustrated with physical examination for detecting a fracture in ankle trauma, using x-ray film as the reference standard

	D: (result of) x-ray		
T: (conclusion of) physical examination	**Fracture**	**No fracture**	**Total**
Fracture	190	80	270
No fracture	10	720	730
Total	200	800	1,000

The SENSITIVITY of test T is the probability of a positive (abnormal) test result in people with disease D: $P(T+|D+) = 190/200 = 0.95$.
The SPECIFICITY of T is the probability of a negative (normal) test result in people without D: $P(T-|D-) = 720/800 = 0.90$.

Note: Sensitivity and specificity together determine the discrimination of a test in a given situation.

The LIKELIHOOD RATIO (LR) of test result T_X is the ratio of probability of test result T_X in people with D, and by the probability of T_X in people without D.

The general formula for LR_X is $\dfrac{P(T_X|D+)}{P(T_X|D-)}$

For a positive result, LR+ is $\dfrac{P(T+|D+)}{P(T+|D-)}$

which is equivalent to $\dfrac{\text{Sensitivity}}{1-\text{specificity}} = \dfrac{190/200}{1-720/800} = 9.5$

For a negative result, LR− is $\dfrac{P(T-|D+)}{P(T-|D-)}$

which is equivalent to $\dfrac{1-\text{Sensitivity}}{\text{specificity}} = \dfrac{1-190/200}{720/800} = 0.06$

Note: LR is an overall measure of the discrimination of test result T_X. The test is useless if LR = 1. The test is better the more LR differs from 1, that is, greater than 1 for LR+ and lower than 1 for LR−.

For tests with multiple outcome categories, LR_X can be calculated for every separate category x, as the ratio of the probability of outcome category x among diseased and the probability of outcome category x among nondiseased.

The PREDICTIVE VALUE of a test result T_X is:
for a positive result, the probability of D in persons with a positive test result:
$P(D+|T+) = 190/270 = 0.70$.
for a negative result, the probability of absence of D in persons with a negative result:
$P(D-|T-) = 720/730 = 0.99$.

Note: The predictive value of a positive test result (posterior or post-test probability) must be compared with the estimated probability of D before T is carried out (the prior or pretest probability). For a good discrimination, the difference between the posttest and the pretest probability should be large.

The (diagnostic) ODDS RATIO (OR), or the cross-product ratio, represents the overall discrimination of a dichotomous test T, and is equivalent to the ratio of LR+ and LR−. OR= $(190 \times 720)/(80 \times 10) = 171$.

Note: If OR=1, T is useless. T is better the more OR differs from 1.

The receiver operating characteristic (ROC) curve graphically represents the relation between sensitivity and specificity for tests with a variable cutoff point, on an ordinal scale (e.g., in case of 5 degrees of suspicion of ankle fracture; or cervical smear) or interval scale (e.g., if degree of suspicion of ankle fracture is expressed in a percentage; or ST changes in exercise ECG testing). If the AUC (area under the curve) = 0.5, the test is useless. For a perfect test, the AUC=1.0 (see Chapter 7).

Table 1.2 Discrimination of some diagnostic tests for various target disorders, expressed in sensitivity, specificity, likelihood ratios, and odds ratio (examples based on several sources)

				Likelihood ratio		
Test	**Target Disorder**	**Sensitivity (%)**	**Specificity (%)**	**Positive result**	**Negative result**	**Odds ratio**
Exercise ECG*[13]	Coronary stenosis	65	89	5.9	0.39	15.0
Absence of chest wall tenderness in chest pain[14]	Myocardial infarction	92	34	1.4	0.23	5.9
Stress thallium scintigraphy[13]	Coronary stenosis	85	85	5.7	0.18	32.1
Ultrasonography[13]	Pancreatic cancer	70	85	4.7	0.35	13.2
CT scan[13]	Pancreatic cancer	85	90	8.5	0.17	51.0
Angiography[13]	Pancreatic cancer	75	80	3.8	0.31	12.0
Gross hematuria in men[15]	Urological cancer	64	99	99	0.51	195
Gross hematuria in women[15]	Urological cancer	47	99	83	0.66	126
ESR $\geq$28 mm/1 h**[16]	Malignancy	78	94	13.0	0.23	56.0
ESR $\geq$28 mm/1 h**[16]	Inflammatory disease	46	95	9.2	0.57	16.2
Intermittent claudication**[17]	Peripheral arterial occlusive disease	31	93	4.4	0.74	5.6
Posterior tibial/dorsalis pedis artery pulse**[17]	Peripheral arterial occlusive disease	73	92	9.1	0.29	30.4
Change in bowel habit**[18]	Colorectal cancer	88	72	3.1	0.17	18.4
Weight loss**[18]	Colorectal cancer	44	85	2.9	0.66	4.6
Rectal bleeding[19]	Colorectal cancer	29	99	68	0.70	97
ESR $\geq$30 mm/1 h**[18]	Colorectal cancer	40	96	10.0	0.42	14.0
White blood cell count .10^9**[18]	Colorectal cancer	75	90	7.5	0.28	26.3
Occult blood test $\geq$1 positive out of 3**[18]	Colorectal cancer	50	82	2.7	0.61	4.6

*Cut-off point: ST depression $\geq$1 mm.

**In a general-practice setting.

- *Monitoring the clinical course* of a disorder, or a state of health such as pregnancy, or the clinical course of an illness during or after treatment.
- *Measuring physical fitness* in relation to specific requirements, for example, for sports or employment.

The evaluation of a diagnostic test concentrates on its added value for the intended application, taking into consideration the burden for the patient (such as pain, cost, or waiting time) and any possible complications resulting from the test (such as intestinal perforation in endoscopy). This requires a comparison between the situations with and without the use of the test or a comparison with the use of other tests.

Prior to the evaluation, one must decide whether to focus on maximizing the health perspectives of the individual patient (which is usually the physician's aim) or on the best possible cost-effectiveness (as economists are likely to do). The latter can be expressed in the amount of money to be invested per number of life years gained, whether or not adjusted for quality of life. Between these two approaches, which do not necessarily yield the same outcome, there is the tension between strictly individual and collective interests. This becomes especially obvious when policy makers have to decide which options would be accepted as the most efficient in a macroeconomic perspective.

Another prior decision is whether one would be satisfied with a qualitative understanding of the diagnostic decision-making process or would need a detailed quantitative analysis.[20] In the first case, one would chart the stages and structure of the decision-making process in relation to the test to be evaluated. This may already provide sufficient insight, for instance, if it becomes clear beforehand that the result will not influence the decision to be taken. Examples of useless testing are (1) the value of the routine electrocardiogram in acute chest pain for exploring the likelihood of a suspected myocardial infarction, with the consequent decision whether to admit the patient to hospital; and (2) the value of "routine blood tests" in general practice for deciding whether to refer a patient with acute abdominal pain to a surgeon. In addition to qualitatively mapping the structure of the decision-making process, quantitative analysis attempts to assess test discrimination and the (probability of the) ultimate clinical outcome, taking the risks (and the costs) of the test procedure into account. The choice of a qualitative or a quantitative approach depends on the question to be answered and the data available.

If a test has not yet been introduced in clinical practice, the prospects for a good evaluation are better than if it is already in general use. It is then still possible to define an appropriate control group the test is not applied to, so that its influence on the prognosis can be investigated. In addition, at such an early stage, the conclusion of the analysis can still be used in the decision regarding the introduction. Furthermore, it is possible to prospectively plan a monitoring procedure and an evaluation after the introduction. All of this emphasizes the importance of developing an evaluation program before a test is introduced.

A common misunderstanding is that only expensive, advanced diagnostic technology cause unacceptable increases in health care costs. Cheap but

frequently used (routine) tests not only account for a major part of direct costs but also greatly influence other costs, as they often preselect patients for more expensive procedures. Yet the performance of such low-threshold diagnostics has often not been adequately evaluated. Examples include many applications of hematological, clinicochemical, and urine tests.[21,22,23]

Methodological challenges

In the evaluation of diagnostic procedures, a number of methodological challenges have to be considered.

Complex relations

Most diagnostics have more than one indication or are relevant for more than one nosological outcome. In addition, tests are often not applied in isolation but in combination, for instance, in the context of clinical protocols. Ideally, clinical research should reflect the health care context,[24] but it is generally impossible to investigate all aspects in one study. One limitation is that the inclusion of a large number of interacting variables calls for large sample sizes that are not easy to obtain. Generally, choices must be made about which issues are the most important. Multivariable statistical techniques are available to allow for the (added) value of various diagnostic data, both separately and in combination, also in the form of diagnostic prediction rules.[25,26,27] In epidemiologic research, such techniques were earlier used for the purpose of analyzing etiologic data, generally focusing on the overall etiologic impact of a factor adjusted for covariables. Diagnostic analysis aims to specify test performance in clinical subgroups or to identify the set of variables that yield the best individual diagnostic prediction, which is a completely different perspective. Much work remains to be done to improve the methodology of diagnostic data analysis.

Diagnostic data analysis will be discussed further in Chapter 7 and multivariable analysis in Chapter 8.

The "gold" standard problem

To evaluate the discriminatory power of a test, its results must be compared with an independently established standard diagnosis. However, a "gold" standard, providing full certainty on the health status, rarely exists. Even x-rays, CT scans, and pathological preparations may produce false positive and false negative results. The aim must then be to define an adequate reference standard that approximates the gold standard as closely as possible.

Sometimes one is faced with whether any appropriate reference standard procedure exists at all. For example, in determining the discrimination of liver tests for diagnosing liver pathology, neither imaging techniques nor biopsies can detect all abnormalities. In addition, as a liver biopsy is an invasive procedure, it is unsuitable for use as a standard in an evaluation study. A useful independent standard diagnosis may not even exist conceptually, for example,

when determining the predictive value of symptoms that are themselves part of the disease definition, as in migraine, or when the symptoms and functionality are more important for management decisions than the anatomical status, as in prostatism. Also, in studying the diagnostic value of clinical examination to detect severe pathology in nonacute abdominal complaints, a comprehensive invasive standard screening, if at all possible or ethically allowed, would yield many irrelevant findings while not all relevant pathology would be immediately found. An option, then, is diagnostic assessment after a follow-up period by an independent panel of experts, representing a "delayed type" cross-sectional study.[28] This may not be perfect but can be the most acceptable solution.[1]

A further issue is the dominance of prevailing reference standards. For example, as long as classic angiography is considered the standard when validating noninvasive vascular imaging techniques, the latter will always seem inferior because perfect agreement is never attainable. However, as soon as a new method comes to be regarded as sufficiently valid to be accepted as the standard, the difference will, from then on, be explained in favor of this new method. In addition, one must accept that two methods may actually measure different concepts. For example, when comparing advanced ultrasound measurements in blood vessels with angiography, the first measures blood flow, relevant to fully explain the symptoms, whereas the second reflects the anatomical situation, which is important for the surgeon. Furthermore, the progress of clinicopathological insights is of great importance. For instance, although clinical pattern X may first be the standard to evaluate the significance of microbiological findings, it will become of secondary diagnostic importance once the infectious agent-causing X has been identified. The agent will then be the diagnostic standard, as illustrated by the history of the diagnosis of tuberculosis.

In Chapters 3 and 6, more will be said about reference standard problems.

Spectrum and selection bias

The evaluation of diagnostics may be flawed by many types of bias.[1,29,30] The most important of these are spectrum bias and selection bias.

Spectrum bias may occur when the discrimination of the diagnostic is assessed in a study population with a different clinical spectrum (for instance, in more advanced cases) than among those in whom the test is to be applied in practice. This may, for example, happen with tests calibrated in a hospital setting but applied in general practice. Also, sensitivity may be determined in seriously diseased subjects, whereas specificity is tested in clearly healthy subjects. Both will then be grossly overestimated relative to the practical situation, where testing is necessary because it is impossible to clinically distinguish in advance who is healthy and who is diseased.

Selection bias is to be expected if there is a relation between the test result and the probability of being included in the study population in which the test is calibrated. For example, subjects with an abnormal exercise

electrocardiogram are relatively likely to be preselected for coronary angiography. Consequently, if this exercise test is calibrated among preselected subjects, a higher sensitivity and a lower specificity will be found than if this preselection had not occurred.[31] Similarly, on the basis of referral patterns alone, it is to be expected that the sensitivity of many tests is higher in the hospital than in general practice and the specificity lower. Considering selection is especially relevant given the nature of medical practice. If a patient enters the consultation room, the physician immediately has information about the patient's gender, age, and general health. The patient can look tired or energetic, visit the physician in his or her office or ask for a house call, or be rushed into the emergency department by ambulance. Such characteristics are in fact results of previous "tests," providing prior information before the physician asks the first question or performs physical examination. This may influence not only the prior probability of disease but also the discrimination of subsequent diagnostic tests.[12]

Although spectrum and selection biases are often related, the clinical picture is the primary point of concern with spectrum bias, whereas the mechanism of selection is the principal issue with selection bias. These types of bias may affect not only sensitivity and specificity but also all other measures of discrimination listed in Table 1.1.[32]

Chapters 2 and 6 will further address the issue of dealing with spectrum and selection phenomena.

"Soft" measures

Subjective factors such as pain, feeling unwell, and reassurance are important in diagnostic management. Most decisions for a watchful waiting strategy in the early phase of an episode of illness are based on the appraisal of such "soft" measures. These often determine the indication for diagnostic examinations and may themselves be part of the diagnostics (e.g., a symptom or complaint) to be evaluated. Also, weighing such factors is generally indispensable in the assessment of the overall clinical outcome and the related impact on quality of life.[33] Evaluation studies should, on the one hand, aim as much as possible to objectify these subjective factors in a reproducible way. On the other hand, interindividual and even intraindividual differences will always play a part[34] and should be acknowledged in the clinical decision-making process.

Observer variability and observer bias

Variability between different observers, as well as for the same observer in reading and interpreting diagnostic data, should not only be acknowledged for "soft" diagnostics such as history taking and physical examination but also for "harder" ones like x-rays, CT and MRI scans, and pathological slides. Even tests not involving any human assessment show inter- and intra-instrument variability. Such variability should be limited if the diagnostic is to produce useful information.

At the same time, researchers should beware of systematic observer bias because of prior knowledge about the subjects examined, especially if subjective factors play a role in determining test results or the reference standard.[35] Clearly, if one wishes to evaluate whether a doctor can accurately diagnose an ankle fracture based on history and clinical examination, it must be certain that the doctor is unaware of an available x-ray result; and a pathologist making an independent final diagnosis should not have previous knowledge about the most likely clinical diagnosis.[36] In such situations, "blinding" is required. A different form of observer bias could occur if the diagnosticians are prejudiced in favor of one of the methods to be compared, as they may unconsciously put greater effort into that technique. A further challenge is that the experience and skill required should be equal for the methods compared, if these are to have a fair chance in the assessment. In this respect, new methods are at risk of being disadvantaged, especially shortly after being introduced.

Discrimination does not mean usefulness

For various reasons, a test with good discrimination does not necessarily influence management.

To begin with, a test may add too little to what is already known clinically to alter management. Furthermore, the physician may take insufficient account of the information provided by the test. This is a complex issue. For instance, studies of the consequences of routine blood testing have shown that in some cases an unaltered diagnosis still led to changes in the considered policy.[16] In a classic study on the clinical impact of upper gastrointestinal endoscopy, a number of changes (23%) in management were made in the absence of a change in diagnosis, whereas in many patients (30%) in which the diagnosis was changed, management was not altered.[37] A test may detect a disorder for which no effective treatment is available. For example, the MRI scan provides refined diagnostic information with regard to various brain conditions for which no therapy is yet in prospect. Finally, as already discussed, supplementary test results are not always relevant for treatment decisions.

For this reason, it is important that studies evaluating diagnostic tests increasingly also investigate the tests' influence on management.[38,39,40]

Indication area and prior probability

Whether a test can effectively detect or exclude a particular disorder is influenced by the prior probability of that disorder. A test is generally not useful if the prior probability is either very low or very high: not only will the result rarely influence patient management, but the risk of, respectively, a false positive or a false negative result is relatively high. In other words, there is an "indication area" for the test between these extremes of prior probability.[2,20] Evaluation of diagnostics should therefore address the issue of whether the test could be particularly useful for certain categories of prior probability. For example, tests with a moderate specificity are not useful for screening in an asymptomatic population (with a low prior probability) because of the high risk of false positive results.

Small steps and large numbers

Compared with therapeutic effectiveness studies, evaluation studies of diagnostic procedures have often neglected the question of whether the sample size is adequate to provide the desired information with a sufficient degree of certainty. A problem is that progress in diagnostic decision making often takes the form of a series of small steps to gain in certainty, rather than one big breakthrough. Evaluating the importance of a small step, however, requires a relatively large study population.

Changes over time and the mosaic of evidence

Innovations in diagnostic technology may proceed at such a speed that a thorough evaluation may take longer than the development of even more advanced techniques. For example, the results of evaluation studies on the clinical impact and cost-effectiveness of the CT scan had not yet fully crystallized when the MRI and PET scans appeared on the scene. So, the results of evaluation studies may already be lagging behind when they appear. Therefore, there is a need for general models (scenarios) for the evaluation of particular (types of) tests and test procedures, whose overall framework is relatively stable and into which information on new tests can be entered by substituting the relevant pieces in the whole mosaic. It may be possible to insert new test opportunities for specific clinical pathways or certain subgroups.[41] The mosaic approach allows for a quick evaluation of the impact of new imaging or DNA techniques with better discrimination on the cost-effectiveness of breast cancer screening, if other pieces of the mosaic (such as treatment efficacy) have not changed. Discrimination can often be relatively rapidly assessed by means of a cross-sectional study, which may avoid new prospective studies. The same can be said for the influence of changes in relevant costs, such as fees for medical treatment or the price of drugs.

Research designs

Various methodological approaches are available to evaluate diagnostic technologies, including original clinical research, on the one hand, and systematically synthesizing the findings of already performed empirical studies and clinical expertise, on the other.

For empirical clinical studies, there is a range of design options. The appropriate study design depends on the research question to be answered (Table 1.3). In diagnostic accuracy studies, the relationship between test result and reference standard has to be assessed cross-sectionally. This can be achieved by a cross-sectional survey (which may also include the delayed type cross-sectional design), but especially in early validation studies other approaches (case–referent or test result–based sampling) can be most efficient. Design options for studying the impact of diagnostic testing on clinical decision making and patient prognosis are the "diagnostic randomized controlled trial" (RCT), which is methodologically the strongest approach, and the before–after study. Also, cohort and case–control designs have been shown to

Table 1.3 Methodological options in diagnostic research in relation to study objectives

Study objective	Methodological options
Clinical studies	
Diagnostic accuracy	Cross-sectional study —survey —case–referent sampling —test result–based sampling
Impact of diagnostic testing on prognosis or management	Randomized controlled trial Cohort study Case–control study Before–after study
Synthesizing findings and expertise	
Synthesizing results of multiple studies	Systematic review Meta-analysis
Evaluation of most effective or cost-effective diagnostic strategy	Clinical decision analysis Cost-effectiveness analysis
Translating findings for practice	Integrating results of the above mentioned approaches Developing clinical prediction rules Expert consensus methods Developing guidelines
Integrating information in clinical practice	ICT support studies Studying diagnostic problem solving Evaluation of implementation in practice

ICT, information and communication technology.

have a place in this context. In Chapter 2, the most important strategic considerations in choosing the appropriate design in diagnostic research will be specifically addressed.

Current knowledge can be synthesized by systematic reviews, meta-analyses, clinical decision analysis, cost-effectiveness studies, and consensus methods, with the ultimate aim of integrating and translating research findings for implementation in practice.

In the following sections, issues of special relevance to diagnostic evaluation studies will be briefly outlined.

Clinical studies

A common type of research is the cross-sectional study, assessing the relationship between diagnostic test results and the presence of particular disorders.[42] This relationship is usually expressed in the measures of discrimination included in Table 1.1. Design options are (1) a survey in an "indicated population," representing subjects in whom the studied test would be considered

in practice; (2) sampling groups with (cases) and without disease (referents) to compare their test distributions; or (3) sampling groups with different test results, between which the occurrence of a disease is compared. It is advisable to include in the evaluation already adopted tests, as this is a direct way to obtain an estimate of the added value of the new test. The cross-sectional study will be dealt with in more detail in Chapter 3.

In an RCT, the experimental group undergoes the diagnostic procedure to be evaluated, while a control group undergoes a different (for example, the usual) or no test. This allows the assessment of not only differences in the percentage of correct diagnoses but also the influence of the evaluated test on management and prognosis. A variant is to apply the diagnostic test to all patients but to disclose its results to the caregivers for a random half of the patients, if ethically justified. This constitutes an ideal placebo procedure for the patient. Not only the (added) value of single tests can be evaluated but also different test strategies and even test-treatment protocols can be compared. Although diagnostic RCTs are not easy to carry out and not always necessary or feasible,[43] several important ones have been carried out already some time ago.[44,45,46,47,48,49] Among the best known are the early trials on the effectiveness of breast cancer screening, which have often linked a standardized management protocol to the screening result.[50,51] The randomized controlled trial in diagnostic research is further discussed in Chapter 4.

If the prognostic value of a test is to be assessed and an RCT is not feasible, its principles can serve as the paradigm in applying other methods, such as the cohort study. The difference from the RCT is that the diagnostic information is not randomly assigned, but a comparison is made between two otherwise composed groups.[52] It has the methodological problem that one can never be sure, especially regarding unknown or unmeasurable covariables, whether the compared groups have similar disease or prognostic spectra to begin with. A method providing relatively rapid results regarding the clinical impact of a test is the case–control study. This is often (although not necessarily) carried out retrospectively, that is, after the course and the final status of the patients are known, in subjects who have been eligible for the diagnostic test to be evaluated. It can be studied whether "indicated subjects" showing an adverse outcome (cases) underwent the diagnostic test more or less frequently than indicated subjects without such outcome (controls). A basic requirement is that the diagnostic must have been available to all involved at the time. Well-known examples are case–control studies on the relationship between mortality from breast cancer and participation in breast cancer screening programs.[53,54] This approach is efficient, although potential bias because of lack of prior comparability of tested and nontested subjects must be considered.

The influence of a diagnostic examination on the physician's management can also be investigated by comparing the intended management policies before and after test results are available. Such before–after comparisons (management impact studies) have their own applications, limitations, and

precautionary measures, as reviewed by Guyatt *et al*.[55] The method has, for example, been applied early in determining the added value of the CT scan and in studying the diagnostic impact of hematological tests in general practice.[56,57] The before–after study design will be outlined in Chapter 5.

Although appropriate inclusion and exclusion criteria for study subjects are as important as in therapeutic research, in diagnostic research using and defining such criteria is less well developed. Appropriate criteria are indispensable in order to focus on the clinical question at issue, the relevant spectrum of clinical severity, the disorders to be evaluated, and the desired degree of selection of the study population (e.g., primary care or referred population).[58]

Another issue that deserves attention is the external (clinical) validity of results of diagnostic studies, as prediction models tend to perform better on data from which they were derived than on data in other, comparable populations.[59,60] In addition, differences in settings often play an important role.[32,61] Deciding whether estimates of test accuracy are generalizable and transferable to other settings depends on an understanding of the possible reasons for variability in test discrimination and calibration across settings, as will be highlighted in Chapter 6.

Synthesizing research findings and clinical expertise

Often the problem is not so much a lack of research findings but the lack of a good summary and systematic processing of those findings. A diagnostic systematic review and meta-analysis of the pooled data of a number of diagnostic studies can synthesize the results of those studies, which provides an overall assessment of the value of diagnostic procedures[62,63] and can also help to identify differences in test accuracy between clinical subgroups. In this way, an overview of the current state of knowledge is obtained within a relatively short time. While until recently making diagnostic systematic reviews faced a methodological backlog compared with systematic reviews of treatment, the decision of the Cochrane Collaboration to include the meta-analysis of studies on diagnostic and screening tests has boosted methods development in this field. As differences in spectrum, setting, and subgroups are quite usual in the application and evaluation of diagnostics, between-study heterogeneity is a frequent phenomenon. This largely complicates the quantitative approach to diagnostic reviews and asks for hierarchical (sometimes also called "multilevel") methods. The methodology of systematically reviewing studies on the accuracy of diagnostic tests is elaborated in Chapter 10.

Another important approach is clinical decision analysis, systematically comparing various diagnostic strategies based on their clinical outcome or cost-effectiveness, supported by probability and decision trees. If good estimates of the discrimination and risks of testing, the occurrence and prognosis of suspected disorders, and the "value" of various clinical outcomes are available, a decision tree can be evaluated quantitatively to identify the clinically optimal or most cost-effective strategy. An important element in the decision analytic approach is the combined analysis of diagnostic and therapeutic

effectiveness. In this context, a qualitative analysis can already be very useful. For example, nowadays, noninvasive techniques show a high level of discrimination in diagnosing carotid stenoses, even in asymptomatic patients. These techniques allow improved patient selection (triage) for the invasive and more hazardous carotid angiography, which is needed to make final decisions regarding surgical intervention. But if surgery has not been proven to influence the prognosis of asymptomatic patients clearly favorably compared with nonsurgical management,[64,65] the decision tree is greatly simplified because it no longer would include either angiography or surgery and maybe not even noninvasive testing.

Decision analysis does not always provide an answer. The problem may be too complex to be summarized in a tree, essential data may be missing, and often a lack of agreement on key assumptions regarding the value of outcomes may occur. Therefore, consensus procedures are often an indispensable step in the translational process from clinical research to guidelines for practice. In these procedures, clinical experts integrate the most recent state of knowledge with their experience to agree on clinical guidelines regarding the preferred approach of a particular medical problem, differentiated for relevant subgroups.[66]

In the context of developing guidelines, clinical prediction rules (CPRs) can be important to aid evidence-based clinical descision making.[67] Chapter 11 will discuss CPRs in relation to the clinical context in which they are used and will review methodological challenges in developing and validating them and in assessing their impact.

Integrating information in clinical practice

To help clinical investigators harvest essential diagnostic research data from clinical databases and to support clinicians in making and in improving diagnostic decisions, medical informatics, and ICT (information and communication technology) innovations are indispensable. Therefore, the issue of implementation of CPRs in relation to ICT will be addressed in Chapter 11, including future developments.

The information processing approaches outlined in the previous section constitute links between research findings and clinical practice and can be applied in combination to support evidence-based medicine. How such input can have optimal impact on the diagnostic decision making of individual doctors is, however, far from simple or straightforward. Therefore, given the growing cognitive requirements of diagnostic techniques, studies to increase our insight in diagnostic problem solving by clinicians is an important part of diagnostic research. This topic is discussed in Chapter 12.

Information from good clinical studies, systematic reviews, and guideline construction is necessary but in many cases not sufficient for improving routine practice. In view of this, during the past decade, implementation research has been strongly developed to face this challenge and to facilitate the steps

from clinical science to patient care. Accordingly, Chapter 13 deals with improving test ordering and its cost-effectiveness in clinical practice.

Concluding remarks

Diagnostic technology assessment would be greatly stimulated if formal standards for the evaluation of diagnostics were to be adopted as a requirement for admittance to the market. Health authorities could initiate assembling panels of experts to promote and monitor the evaluation of both new and traditional diagnostic facilities as to effectiveness, efficiency, and safety, as a basis for acceptance and retention of diagnostics in clinical practice. Furthermore, professional organizations have a great responsibility to set, to implement, to maintain, and to improve clinical standards. More effective international cooperation would be useful, as it has proved to be in the approval and quality control of drugs. In this way, the availability of resources for industrial, private, and governmental funding for diagnostic research and technology assessment would also be stimulated.

Regarding the feasibility of diagnostic evaluation studies, the required size and duration must be considered in relation to the speed of technological progress. This speed can be very great, for instance, in areas where molecular genetic knowledge and information and communication technology play an important part. Especially in such areas, updating of decision analyses, expert assessments, and scenarios by inserting new pieces of the "mosaic" of evidence may be more useful than fully comprehensive, lengthy trials. This may be, for example, relevant for the evaluation of diagnostic areas where current tests will be replaced by new DNA diagnostics in the years to come.

References

1. Feinstein AR. *Clinical epidemiology: The architecture of clinical research.* Philadelphia: W. B. Saunders; 1985.
2. Sackett DL, Haynes RB, Tugwell P. *Clinical epidemiology: A basic science for clinical medicine.* Boston: Little, Brown; 1985.
3. Pocock SJ. *Clinical trials, a practical approach.* New York: John Wiley & Sons; 1983.
4. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic research, principles, and quantitative methods.* Belmont (CA): Wadsworth; 1982.
5. Miettinen OS. *Theoretical epidemiology, principles of occurrence research in medicine.* New York: John Wiley & Sons; 1985.
6. Sheps SB, Schechter MT. The assessment of diagnostic tests: A survey of current medical research. *JAMA*. 1984;**252**:2418–22.
7. Reid ML, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic research: Getting better but still not good. *JAMA*. 1995;**274**:645–51.
8. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;**282**:1061–66.
9. Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ*. 2002; **324**:539–41.

10. Bossuyt PM, Reitsma JB, Bruns DE, et al. The Standards for Reporting of Diagnostic Accuracy group: Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Clin Radiol*. 2003;**58**:575–80.
11. Whiting PF, Weswood ME, Rutjes AW, et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol*. 2006;**6**:9.
12. Buntinx F, Knockaert D, de Blaey N, et al. Chest pain in general practice or in the hospital emergency department: Is it the same? *Fam Pract*. 2001;**18**:591–94.
13. Panzer RJ, Black ER, Griner PF, eds. *Diagnostic strategies for common medical problems.* Philadelphia: American College of Physicians; 1991. 2nd ed. Black ER, Bordley DR, Tape TG, Panzer RJ, eds.; 1999.
14. Bruyninckx R, Aertgeerts B, Buntinx F. Signs and symptoms in diagnosing acute myocardial infarction and acute coronary syndrome: A diagnostic meta-analysis. Report. 2007: University of Leuven, 2007.
15. Bruyninckx R, Buntinx F, Aertgeerts B, et al. The diagnostic value of macroscopic haematuria for the diagnosis of urological cancer in general practice. *Br J Gen Pract*. 2003;**53**:31–35.
16. Dinant GJ, Knottnerus JA, Van Wersch JW. Discriminating ability of the erythrocyte sedimentation rate: A prospective study in general practice. *Br J Gen Pract*. 1991;**41**:365–70.
17. Stoffers HEJH, Kester ADM, Kaiser V, et al. Diagnostic value of signs and symptoms associated with peripheral arterial obstructive disease seen in general practice: A multivariable approach. *Med Decis Making*. 1997;**17**:61–70.
18. Fijten GHF. *Rectal bleeding, a danger signal?* Amsterdam: Thesis Publishers; 1993.
19. Wouters H, Van Casteren V, Buntinx F. Rectal bleeding and colorectal cancer in general practice: diagnostic study. *BMJ*. 2000;**321**:998–99.
20. Knottnerus JA, Winkens R. Screening and diagnostic tests. In: Silagy C, Haines A, eds. *Evidence based practice in primary care.* London: BMJ Books; 1998.
21. Dinant GJ. *Diagnostic value of the erythrocyte sedimentation rate in general practice*. PhD thesis, University of Maastricht; 1991.
22. Hobbs FD, Delaney BC, Fitzmaurice DA, et al. A review of near patient testing in primary care. *Health Technol Assess*. 1997;**1**:i–iv, 1–229.
23. Campens D, Buntinx F. Selecting the best renal function tests: A meta-analysis of diagnostic studies. *Int J Technol Assess Health Care*. 1997;**13**:343–56.
24. van Weel C, Knottnerus JA. Evidence-based interventions and comprehensive treatment. *Lancet*. 1999;**353**:916–8.
25. Spiegelhalter DJ, Crean GP, Holden R, et al. Taking a calculated risk: predictive scoring systems in dyspepsia. *Scand J Gastroenterol*. 1987;**22** Suppl 128:152–60.
26. Knottnerus JA. Diagnostic prediction rules: Principles, requirements, and pitfalls. *Prim Care*. 1995;**22**:341–63.
27. Knottnerus JA. Application of logistic regression to the analysis of diagnostic data. *Med Decis Making*. 1992;**12**:93–108.
28. Knottnerus JA, Dinant GJ. Medicine based evidence, a prerequisite for evidence based medicine. *BMJ*. 1997;**315**:1109–10.
29. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;**299**:926–30.
30. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987;**6**:411–23.
31. Green MS. The effect of validation group bias on screening tests for coronary artery disease. *Stat Med*. 1985;**4**:53–61.

32. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. *J Clin Epidemiol*. 1992;**45**:1143–54.
33. Feinstein, AR. *Clinimetrics*. New Haven (CT): Yale University Press; 1987.
34. Zarin OA, Pauker SG. Decision analysis as a basis for medical decision making: The tree of Hippokrates. *J Med Philos*. 1984;**9**:181–213.
35. Moons KG, Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic studies? *J Clin Epidemiol*. 2002;**55**:633–36.
36. Schwartz WB, Wolfe HJ, Pauker SG. Pathology and probabilities, a new approach to interpreting and reporting biopsies. *N Engl J Med*. 1981;**305**:917–23.
37. Liechtenstein JI, Feinstein AR, Suzio KD, et al. The effectiveness of pandendoscopy on diagnostic and therapeutic decisions about chronic abdominal pain. *J Clin Gastroenterol*. 1980;**2**:31–36.
38. Sachs S, Bilfinger TV. The impact of positron emission tomography on clinical decision making in a university-based multidisciplinary lung cancer practice. *Chest*. 2005;**128**(2):698–703.
39. Milas M, Stephen A, Berber E, et al. Ultrasonography for the endocrine surgeon: a valuable clinical tool that enhances diagnostic and therapeutic outcomes. *Surgery*. 2005;**138**:1193–200.
40. Kymes SM, Lee K, Fletcher JW, et al. Assessing diagnostic accuracy and the clinical value of positron emission tomography imaging in patients with solitary pulmonary nodules (SNAP). *Clin Trials*. 2006;**3**:31–42.
41. Bossuyt PM, Irwig L, Craig J, et al. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ*. 2006;**332**:1089–92.
42. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. *J Clin Epidemiol*. 2003;**56**:1118–28.
43. Guyatt GH, Sackett DL, Haynes RB. Evaluating diagnostic tests. In: Haynes RB, Sackett DL, Guyatt GH, Tugwell P, eds. *Clinical epidemiology: How to do clinical practice research*. Philadelphia: Lippincott, Williams & Wilkins; 2006.
44. Dronfield MW, Langman MJ, Atkinson M, et al. Outcome of endoscopy and barium radiography for acute upper gastrointestinal bleeding: controlled trial in 1037 patients. *BMJ*. 1982;**284**:545–48.
45. Brett GZ. The value of lung cancer detection by six-monthly chest radiographs. *Thorax*. 1968;**23**:414–20.
46. Brown VA, Sawers RS, Parsons RJ, et al. The value of antenatal cardiotocography in the management of high risk pregnancy: a randomised controlled trial. *Br J Obstet Gynaecol*. 1982;**89**:716–22.
47. Flynn AM, Kelly J, Mansfield H, et al. A randomised controlled trial of non-stress antepartum cardiotocography. *Br J Obstet Gynaecol*. 1982;**89**:427–33.
48. Durbridge TC, Edwards F, Edwards RG, et al. An evaluation of multiphasic screening on admission to hospital. *Med J Aust*. 1976;**1**:703–5.
49. Hull RD, Hirsch J, Carter CJ, et al. Diagnostic efficacy of impedance phletysmography for clinically suspected deep-vein thrombosis: A randomized trial. *Ann Intern Med*. 1985;**102**:21–28.
50. Shapiro S, Venet W, Strax Ph, et al. Ten- to fourteen-year effect of screening on breast cancer mortality. *J Natl Cancer Inst*. 1982;**69**:349–55.
51. Tabár L, Fagerberg CJG, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography. *Lancet*. 1985;**1**:829–31.
52. Harms LM, Schellevis FG, van Eijk JT, et al. Cardiovascular morbidity and mortality among hypertensive patients in general practice: the evaluation of long-term systematic management. *J Clin Epidemiol*. 1997;**50**:779–86.

53. Collette HJA, Day NE, Rombach JJ, et al. Evaluation of screening for breast cancer in a non-randomised study (the DOM project) by means of a case control study. *Lancet*. 1984;**1**:1224–6.
54. Verbeek ALM, Hendriks JHCL, Holland R, et al. Reduction of breast cancer mortality through mass-screening with modern mammography. *Lancet*. 1984;**1**:1222–4.
55. Guyatt GH, Tugwell P, Feeny DH, et al. The role of before–after studies of therapeutic impact in the evaluation of diagnostic technologies. *J Chronic Dis*. 1986;**39**:295–304.
56. Fineberg HV, Bauman R. Computerized cranial tomography: effect on diagnostic and therapeutic planns. *JAMA*. 1977;**238**:224–7.
57. Dinant GJ, Knottnerus JA, van Wersch JW. Diagnostic impact of the erythrocyte sedimentation rate in general practice: A before–after analysis. *Fam Pract*. 1991;**9**:28–31.
58. Knottnerus JA. Medical decision making by general practitioners and specialists. *Fam Pract*. 1991;**8**:305–7.
59. Starmans R, Muris J, Schouten HJA, et al. The diagnostic value of scoring models for organic and non-organic gastrointestinal disease in non-acute abdominal pain. *Med Decis Making*. 1994;**14**:208–15.
60. Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol*. 2003;**56**:441–7.
61. Irwig L, Bossuyt P, Glasziou P, et al. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ*. 2002;**324**:669–71.
62. Irwig L, Macaskill P, Glasziou P, et al. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol*. 1995;**48**:119–30.
63. Buntinx F, Brouwers M. Relation between sampling device and detection of abnormality in cervical smears: A meta-analysis of randomised and quasi-randomised studies. *BMJ*. 1996;**313**:1285–90.
64. Chambers BR, Donnan GA. Carotid endarterectomy for asymptomatic carotid stenosis. *Cochrane Database Syst Rev*. 2005;Oct 19(4):CD001923.
65. Ederle J, Brown MM.The evidence for medicine versus surgery for carotid stenosis. *Eur J Radiol*. 2006;60:3–7.
66. Woolf SH, Grol R, Hutchinson A, et al. Clinical guidelines: Potential benefits, limitations, and harms of clinical guidelines. *BMJ*. 1999;**318**:527–30.
67. McGinn T, Guyatt G, Wyer P, et al. Diagnosis: Clinical prediction rules: Users' guides to the medical literature. Chicago: AMA Press, 2004. pp. 471–83.